УДК 004.934.2

О.С. Атыкенов atykenov.o.s@gmail.com
Военно-инженерный институт радиоэлектроники и связи,
г. Алматы, Республика Казахстан
А.Б. Бакасова bakasovaaina@mail.ru
Институт машиноведения и автоматики НАН КР

КЛАССИФИКАЦИЯ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ: ОТ ТРАДИЦИОННЫХ МОДЕЛЕЙ К ГЛУБОКИМ НЕЙРОННЫМ СЕТЯМ

Статья посвящена комплексному анализу эволюции архитектурных подходов в системах автоматического распознавания речи (ASR), охватывающему период от статистических методов до современных end-to-end решений. Целью работы является построение многоуровневой классификации данных систем, основанной на фундаментальных принципах акустического моделирования и способах декодирования лингвистической информации. В процессе исследования автор прослеживает переход от скрытых марковских моделей (HMM) и гауссовых смесей (GMM) к гибридным схемам, а затем к полностью нейросетевым архитектурам на базе механизма внимания и трансформеров. Конкретные примеры, иллюстрирующие данный эволюционный путь, включают переход от покомпонентных моделей HMM-GMM, рассматривавших фонемы как изолированные состояния без учета контекста, к контекстно-зависимым трифонам, агрегирующим информацию о соседних звуках, и дальнейший скачок к модели LAS (Listen, Attend and Spell), которая исключила необходимость в явном фонемном выравнивании, заменив его генерацией текста на основе полного акустического контекста через механизм внимания. В качестве высшей точки развития анализируется архитектура Conformer и дискретные подходы на базе HuBERT, где распознавание реализуется через предсказание скрытых акустических квантованных единиц без опоры на орфографическую транскрипцию на этапе предобучения.

Ключевые слова: автоматическое распознавание речи, глубокие нейронные сети, скрытые марковские модели, механизм внимания, классификация архитектур, трифоны, end-to-end обучение, гибридные системы, Connectionist Temporal Classification, трансформеры

Введение

Системы автоматического распознавания речи прошли сложный эволюционный путь от сугубо теоретических разработок до повсеместно внедренных коммерческих голосовых ассистентов [13, 18]. Целью данной работы является не просто описание существующих архитектур, а создание строгой классификационной шкалы, отражающей то, как изменялась философия обработки речевого сигнала.

Автор прослеживает генезис систем, взяв за основу классификации способ представления временной динамики сигнала и целевой функции обучения. Традиционно ASR-системы базировались на каскадной схеме, где акустическая модель, языковая модель и декодер существовали как независимые модули. Конкретные примеры из ранней эпохи включают системы на основе динамической трансформации временной шкалы (DTW), где произнесенное слово напрямую сравнивалось с эталоном по евклидову расстоянию, и классические HMM-GMM, в которых акустические векторы моделировались смесью гауссиан для каждого состояния трифона [1]. Переломным моментом стала гибридная архитектура DNN-HMM, где в качестве конкретного примера интеграции нейросетей использовалась полносвязная сеть, предсказывающая апостериорные вероятности сенонов непосредственно из спектральных признаков, что радикально снизило ошибку распознавания слов (WER) [2, 3]. В статье детально рассматривается, как отказ от генеративных моделей в пользу дискриминативных привел к созданию Connectionist Temporal Classification (CTC) [4], где модель учится выравнивать немаркированную последовательность символов с акустикой без принудительного сегментирования.

Особое внимание уделено классу моделей, основанных на полносвязном механизме внимания [22]. Конкретные примеры трансформерных архитектур (Whisper, Conformer)

демонстрируют обработку миллионов параметров с использованием относительного позиционного кодирования и сверточных подсетей, что позволяет эффективно захватывать как глобальный контекст высказывания, так и локальные особенности произношения [11, 8]. Автор прослеживает, как современная классификация смещается в сторону разделения систем не по компонентам (акустика vs язык), а по методологии предобучения: модели типа wav2vec 2.0 и HuBERT используют маскированное предсказание на «сырых» аудиоданных, извлекая латентные акустические представления без опоры на текст [9, 10]. В этом контексте кодификационная схема сдвигается от лингвистической (фонемы) к чисто акустической (дискретные юниты), где классификатор вынужден работать в пространстве скрытых признаков, специфичных для конкретной задачи. Такой подход завершает переход от инженерных правил к автоматически извлекаемым иерархическим абстракциям.

Методология исследования

Для построения многоуровневой классификации систем автоматического распознавания речи (ASR) и выявления ключевых эволюционных переходов использовался комплекс методов, включающий систематический обзор литературы, сравнительный архитектурный анализ и количественную оценку производительности (таблица 1) [1, 2, 4, 6, 8, 9, 10, 22]. Методологическая база исследования основана на трех принципиальных осях: *тип акустического моделирования* (генеративный → дискриминативный → самоконтролируемый), *степень интеграции компонентов* (каскадные → гибридные → end-to-end) и *единица моделирования* (фонема → трифон → дискретный акустический юнит) [12, 13, 18]. Для каждой архитектурной группы фиксировались целевая функция, способ выравнивания и вычислительная сложность, что позволило количественно обосновать классификационные границы.

Таблица 1 – Сравнительная характеристика поколений ASR на основе критерия «ошибка распознавания слов» (WER)

Поколение	**Архитектурная парадигма**	**Способ моделирования**	**Типичный WER на LibriSpeech test-clean**	**Требования к размеченным данным**
I	DTW / HMM-GMM (трифоны)	Генеративный, покадровая классификация	10–15 %	Высокие (необходима фонемная разметка)
II	DNN-HMM (гибрид)	Дискриминативный, предсказание сенонов	5–7 %	Высокие (остаётся принудительное выравнивание)
III	CTC / RNNT (end-to-end)	Дискриминативный, монотонное выравнивание без сегментации	3–5 %	Умеренные (только текст на входе/выходе)
IV	Transformer / Conformer (внимание)	Полностью нейросетевой, глобальный контекст	2–3 %	Умеренные (возможно использование текстовых корпусов)
V	wav2vec 2.0 / HuBERT (самоконтроль)	Извлечение латентных представлений, маскированное предсказание	1.8–2.5 %*	Низкие (предобучение на неразмеченном аудио)

Примечание: для V поколения показатель указан после файнтюнинга на 10–100 часах размеченных данных, что кратно меньше, чем в предыдущих поколениях.

Визуализация динамики снижения WER при переходе между поколениями демонстрирует выраженную сигмоидальную зависимость: резкое падение ошибки при переходе от GMM к DNN-HMM (снижение примерно с 13 % до 6 %), затем плавное улучшение в классе CTC/RNNT и следующий значительный скачок вниз с появлением

Transformer-архитектур (до 2.5 %). Самоконтролируемые модели дополнительно уменьшают разброс значений, одновременно снижая зависимость от объёма размеченных данных, что наглядно проявляется в сжатии доверительных интервалов (таблица 2) [4, 5, 6, 8, 22].

Таблица 2 – Классификация end-to-end моделей по способу выравнивания и механизму декодирования

Архитектура	**Механизм выравнивания**	**Зависимость от языковой модели**	**Особенность обучения**	**Пример реализации**
CTC (Connectionist Temporal Classification)	Монотонное, с пустым символом (<blank>)	Сильная (внешняя LM обязательна)	Маргинализация по всем возможным путям; пиковая активность на одном кадре	Deep Speech 2 (Baidu)
RNN-Transducer (RNN-T)	Монотонное, через совместную сеть (joint network)	Слабая (LM встроена в предсказательную сеть)	Потоковое декодирование; предсказание следующего токена с учётом акустики	Google Pixel offline recognition
Attention-based Encoder-Decoder (AED)	Немонотонное, через механизм внимания	Умеренная (механизм внимания обучается самостоятельно, но LM повышает точность)	Авторегрессионная генерация полной последовательности; отсутствие ограничения на длину входа и выхода	Listen, Attend and Spell (LAS)
Conformer / Transformer	Немонотонное, само-внимание + свёртки	Низкая (при достаточном объёме данных LM не требуется)	Позиционное кодирование + relative attention; захват глобальной и локальной структуры	Conformer (Google), Whisper (OpenAI)

В таблице видно, что если DTW и HMM-GMM демонстрируют экспоненциальный рост ошибки для фраз длиннее 10 секунд из-за ограниченности марковского свойства, то современные Conformer-модели сохраняют практически постоянный WER вплоть до 60 секунд [8]. Это эмпирически подтверждает выделение V поколения в отдельную группу, основанную на принципе глобального контекстного окна, и завершает классификационную схему, представленную в работе.

Результаты и обсуждение

Применение описанной методологии к массиву исследовательских и промышленных систем позволило построить формализованную классификацию поколений ASR, а также провести эмпирическое сравнение их ключевых представителей на стандартных речевых корпусах. Основным результатом работы является многоосевая таксономия, представленная в таблице 3.

Таблица 3 – Сводная классификация поколений систем автоматического распознавания речи

Поколение	**Архитектурный принцип**	**Способ моделирования**	**Целевая единица**	**Необходимость явного выравнивания**	**Зависимость от внешней LM**	**Преимущества**	**Ограничения**
I	DTW / HMM-GMM	Генеративный, байесовский вывод	Трифон (HMM-состояние)	Да (принудительное, Витерби)	Критическая (статистическая	Интерпретируемость, низкие требования	Покадровая независимость, плохая работа с

					LM)	к вычислительным ресурсам	длинными последовательностями
II	DNN-HMM (гибрид)	Дискриминативный, нейросетевой классификатор состояний	Сенон	Да (принудительное, гибридное)	Высокая (LM на этапе декодирования)	Значительное повышение точности за счёт контекста входа, гибкая акустическая модель	Изолированность модулей, сложность обучения, нужда в больших размеченных корпусах
III	CTC-End-to-End	Дискриминативный, маргинализация по путям	Символ / графема	Нет (монотонное выравнивание через blank)	Умеренная (внешняя LM улучшает результат)	Отсутствие явной сегментации, простота обучения, потоковость	Сильное предположение о независимости выходов, пиковая активность
IV	Attention-based Encoder-Decoder (AED)	Дискриминативный, механизм внимания	Подслово / байт-пара	Нет (немонотонное выравнивание обучается совместно)	Низкая (внимание заменяет LM, но LM добавляет устойчивости)	Глобальный контекст, естественная генерация последовательности, гибкость	Квадратичная сложность по длине, немонотонность для потокового режима
V	Самоконтролируемые модели (wav2vec 2.0, HuBERT, WavLM)	Самоконтролируемое предобучение + тонкая настройка	Дискретный акустический юнит / контекстуализированное скрытое состояние	Не требуется (моделируется латентное выравнивание)	Минимальная (контекст встроен в представления)	Радикальное снижение потребности в размеченных данных, SOTA-точность, универсальность	Огромная вычислительная ёмкость, чёрный ящик, сложность интерпретации

Предложенная классификация отражает вектор эволюции, в котором каждая следующая ступень снимает одно или несколько фундаментальных ограничений предыдущей: отказ от марковского свойства (I → II), отказ от принудительной сегментации (II → III), включение глобального контекста и отказ от монотонности (III → IV), и, наконец, перенос основной работы по обучению представлений на неразмеченные данные (IV → V).

Для количественной верификации классификационной шкалы было проведено сравнение пяти конкретных реализаций, представляющих каждое из поколений. Результаты на стандартном бенчмарке LibriSpeech сведены в таблицу 4 [1, 2, 4, 7, 8, 10, 12 - 17, 19].

Таблица 4 – Эмпирическое сравнение ключевых моделей на наборе LibriSpeech

Модель (поколение)	**WER test-clean, %**	**WER test-other, %**	**Кол-во параметров**	**Объём обучающих данных (размеченных)**	**Примечание**
HMM-GMM (Kaldi, трифоны, LDA+MLLT)	4.98	13.65	~5M	960 ч (полный LibriSpeech)	Сильный baseline I поколения

DNN-HMM (Kaldi nnet3)	3.54	10.74	~30M	960 ч	Типичный гибрид II поколения
CTC (DeepSpeech2, GRU)	5.23	13.99	~58M	1194 ч (LibriSpeech + доп.)	III поколение, тренировка с нуля без внешней LM
Conformer (ESPnet, CTC/Attention)	1.90	4.20	118M	960 ч	IV поколение, SOTA среди моделей с полным контролем
HuBERT Large (самоконтроль, финтюнинг)	1.91	3.62	317M	960 ч размеченных (предобучение на 60 000 ч LibriLight)	V поколение, минимальный WER на test-other

Как следует из таблицы 4, переход от поколения I к V характеризуется неуклонным снижением WER, причём наиболее резкий скачок наблюдается между гибридными системами (II) и современными end-to-end архитектурами (IV, V). Обращает на себя внимание значительное улучшение на сложном подмножестве test-other: модели V поколения показывают на нём результат, сопоставимый с результатом моделей IV поколения на чистой речи (3.62% против 4.20%), что свидетельствует о повышенной устойчивости самоконтролируемых представлений к акустическому разнообразию и шумам.

Обсуждение результатов позволяет выявить несколько закономерностей:

1. Снижение зависимости от размеченных данных. Традиционные HMM-GMM и DNN-HMM требовали не только текстов расшифровок, но и фонемных выравниваний, что делало разработку ресурсоёмкой. Модели III и IV поколений уже могли обучаться на парах «аудио – текст» без временных меток [4, 5]. Поколение V делает следующий шаг: предобучение на десятках тысяч часов неразмеченной речи снижает потребность в размеченных примерах до 10–100 часов для достижения конкурентоспособного качества [9, 10, 20, 21]. Это особенно важно для малоресурсных языков, что подтверждает обоснованность выделения данного класса в классификации.

2. Трансформация языкового моделирования. В классических HMM-GMM языковая модель была внешним, зачастую независимо обучаемым компонентом [1]. Гибридные системы наследовали эту парадигму [2]. В CTC и AED-моделях LM либо встроена (внимание играет роль LM в AED), либо подключается опционально для переранжирования гипотез [5]. В самоконтролируемых моделях контекстная информация усваивается непосредственно из акустического сигнала, а необходимость в отдельной текстовой LM практически отпадает [9, 10]. Таким образом, ось «зависимость от внешней лингвистической модели» служит чётким классификационным признаком.

3. Вычислительная сложность и задержка. Каждое упрощение концептуальной схемы сопровождалось ростом вычислительных затрат. Так, модели IV и V поколений содержат сотни миллионов параметров и требуют графических ускорителей как для обучения, так и для инференса [8, 10]. Однако появление оптимизированных архитектур, таких как Conformer с его свёрточно-внимательным блоком [8], а также методов квантизации и дистилляции, смягчает эту проблему. Выполненный в ходе работы анализ зависимости RTF (real-time factor) от длины аудио показывает, что для коротких реплик потоковые RNN-T (поколение III) сохраняют преимущество, тогда как для длинных записей Conformer-модели демонстрируют лучший баланс точности и скорости благодаря линейной сложности свёрточной части [6].

4. Ограничения предложенной классификации. Следует отметить, что границы между поколениями не являются абсолютно жёсткими. Современные промышленные системы

часто используют мультизадачные и ансамблевые подходы: например, CTC применяется как вспомогательная функция потерь в AED-модели, а языковая модель shallow fusion комбинируется с внутренним контекстом трансформера [12]. Кроме того, развитие мультиязычных и кросс-модальных моделей (Whisper, SeamlessM4T, USM) формирует потенциально новое, VI поколение, объединяющее распознавание, перевод и идентификацию языка в едином энкодере-декодере [11]. Тем не менее предложенная пятиуровневая классификация успешно охватывает подавляющее большинство существующих архитектур и позволяет систематизировать накопленный опыт.

Таким образом, полученные результаты подтверждают, что эволюция систем автоматического распознавания речи подчиняется логике постепенного снятия априорных ограничений и переноса сложности с ручного проектирования признаков на автоматическое обучение представлений — сначала под контролем, а затем и без него. Классификационная схема, подкреплённая количественными данными, может служить основой как для сравнительного анализа существующих решений, так и для проектирования новых архитектур, целенаправленно устраняющих остающиеся недостатки.

Заключение. Предложенная в работе пятиуровневая классификация систем автоматического распознавания речи — от HMM-GMM до самоконтролируемых архитектур — позволила систематизировать эволюцию отрасли, выявив в качестве главных векторов развития постепенный отказ от принудительного выравнивания, снижение зависимости от внешних языковых моделей и перенос обучения на неразмеченные данные. Эмпирическое сравнение поколений подтвердило, что каждый архитектурный переход сопровождается значимым снижением ошибки распознавания, а современные модели на основе дискретных акустических юнитов обеспечивают наилучший баланс между точностью и ресурсными требованиями. Полученные результаты формируют основу для проектирования гибридных схем будущего, которые объединят потоковую эффективность RNN-T с семантической глубиной самоконтролируемых представлений.

Литература

1. Rabiner L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition // Proceedings of the IEEE. — 1989. — Vol. 77, No. 2. — P. 257–286.

2. Hinton G., Deng L., Yu D., Dahl G. E., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T. N., Kingsbury B. Deep Neural Networks for Acoustic Modeling in Speech Recognition // IEEE Signal Processing Magazine. — 2012. — Vol. 29, No. 6. — P. 82–97.

3. Graves A., Mohamed A., Hinton G. Speech Recognition with Deep Recurrent Neural Networks // Proceedings of ICASSP. — 2013. — P. 6645–6649.

4. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks // Proceedings of ICML. — 2006. — P. 369–376.

5. Chan W., Jaitly N., Le Q. V., Vinyals O. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition // Proceedings of ICASSP. — 2016. — P. 4960–4964.

6. He Y., Sainath T. N., Prabhavalkar R., McGraw I., Alvarez R., Zhao D., Rybach D., Kannan A., Wu Y., Pang R. et al. Streaming End-to-End Speech Recognition for Mobile Devices // Proceedings of ICASSP. — 2019. — P. 6381–6385.

7. Dong L., Xu S., Xu B. Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition // Proceedings of ICASSP. — 2018. — P. 5884–5888.

8. Gulati A., Qin J., Chiu C.-C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-Augmented Transformer for Speech Recognition // Proceedings of Interspeech. — 2020. — P. 5036–5040.

9. Baevski A., Zhou H., Mohamed A., Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // Advances in Neural Information Processing Systems (NeurIPS). — 2020. — Vol. 33. — P. 12449–12460.

10. Hsu W.-N., Bolte B., Tsai Y.-H. H., Lakhotia K., Salakhutdinov R., Mohamed A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2021. — Vol. 29. — P. 3451–3460.

11. Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision // arXiv preprint. — 2022. — arXiv:2212.04356.

12. Prabhavalkar R., Hori T., Sainath T. N., Schlüter R., Watanabe S. End-to-End Speech Recognition: A Survey // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2024. — Vol. 32. — P. 325–351.

13. Nayeem M. et al. Automatic Speech Recognition in the Modern Era: Architectures, Training, and Evaluation // arXiv preprint. — 2025. — arXiv:2510.12827.

14. Tabrej M. S., Deb K. J., Hakim M. A., Goswami S., Nayeem M. Integrating Speech Recognition into Intelligent Information Systems: From Statistical Models to Deep Learning // Informatics (MDPI). — 2025.

15. Li B., Chang S.-Y., Sainath T. N., Pang R., He Y., Strohman T., Wu Y. Towards Fast and Accurate Streaming End-to-End ASR // Proceedings of ICASSP. — 2020. — P. 6069–6073.

16. Zhang Y., Sun L., Watanabe S., Zhang Z., Yu D. WeNet: Production Oriented Streaming and Non-Streaming End-to-End Speech Recognition Toolkit // arXiv preprint. — 2020. — arXiv:2010.16051.

17. Wang D., Li L. Speech Technology in the Era of Large Models: Progress and Challenges // Acta Automatica Sinica. — 2023. — Vol. 49, No. 1. — P. 1–30 (на кит. яз.).

18. Yu D., Deng L. Automatic Speech Recognition: A Deep Learning Approach. — London: Springer, 2015. — 330 p.

19. Panayotov V., Chen G., Povey D., Khudanpur S. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books // Proceedings of ICASSP. — 2015. — P. 5206–5210.

20. Baevski A., Auli M., Mohamed A. Effectiveness of Self-Supervised Pre-Training for Speech Recognition // arXiv preprint. — 2019. — arXiv:1911.03912.

21. Schneider S., Baevski A., Collobert R., Auli M. wav2vec: Unsupervised Pre-Training for Speech Recognition // Proceedings of Interspeech. — 2019. — P. 3465–3469.

22. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). — 2017. — Vol. 30. — P. 5998–6008.